Apache Science Data Analytics Platform

Edward M. Armstrong[1], Mark A. Bourassa[2], Tom Cram[3], Jocelyn Elya[2], **Thomas Huang[1]**, Joseph Jacob[1], Zaihua Ji[3],Yongyao Jiang[4], Yun Li[4], Lewis McGibbney[1], Nga Quach[1], Shawn Smith[2], Vardis Tsontos[1], Brian Wilson[1], Steve J .Worley[3], and Chaowei (Phil) Yang[4]

[1] NASA Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Dr., Pasadena, CA 91109, USA
[2] Center for Ocean-Atmospheric Prediction Studies, 2000 Levy Avenue, Building A, Suite 292, Tallahassee, FL 32306-2741, USA
[3] National Center for Atmospheric Research, P.O. Box 3000, Boulder, CO 80307-3000, USA
[4] George Mason University, 4400 University Drive, Fairfax, VA 22030, USA

National Aeronautics and
Space Administration

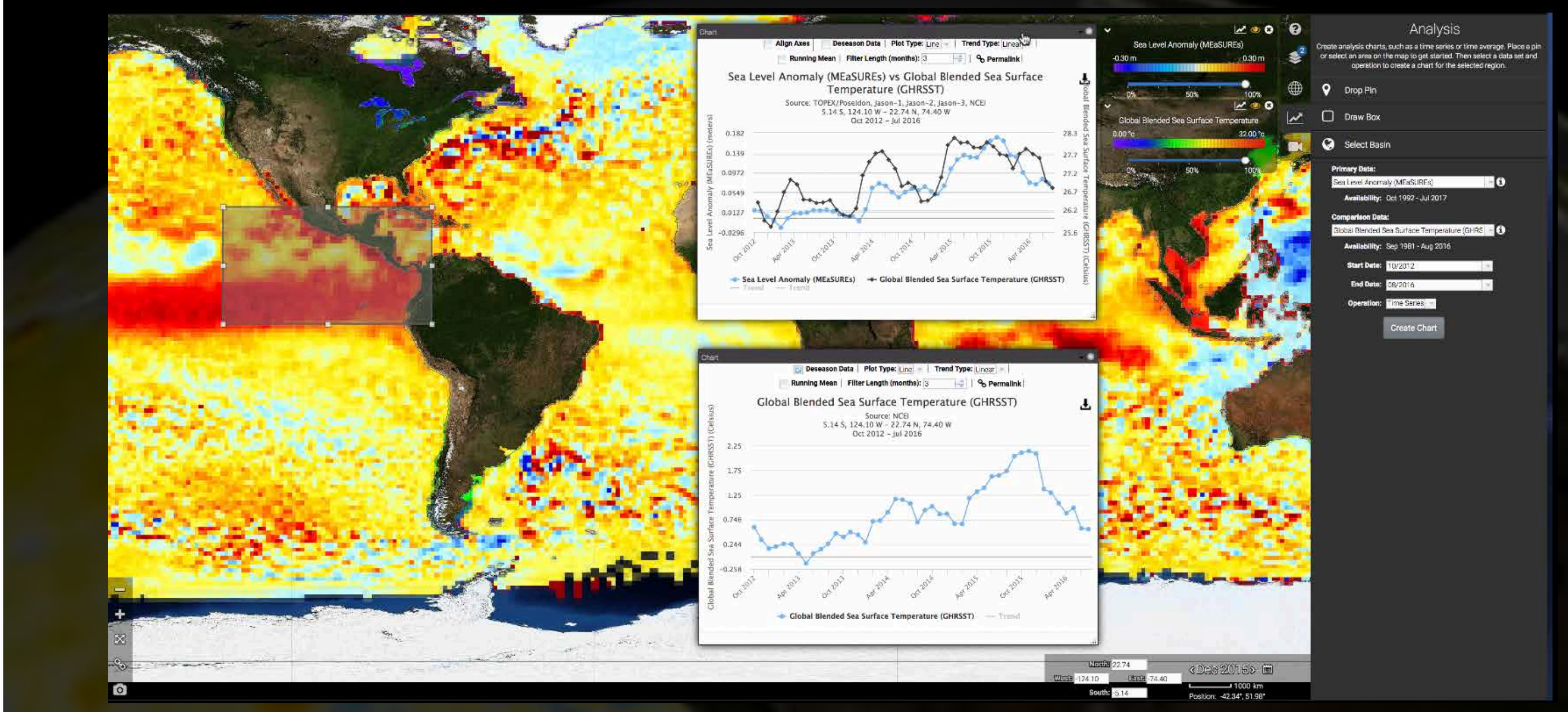
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

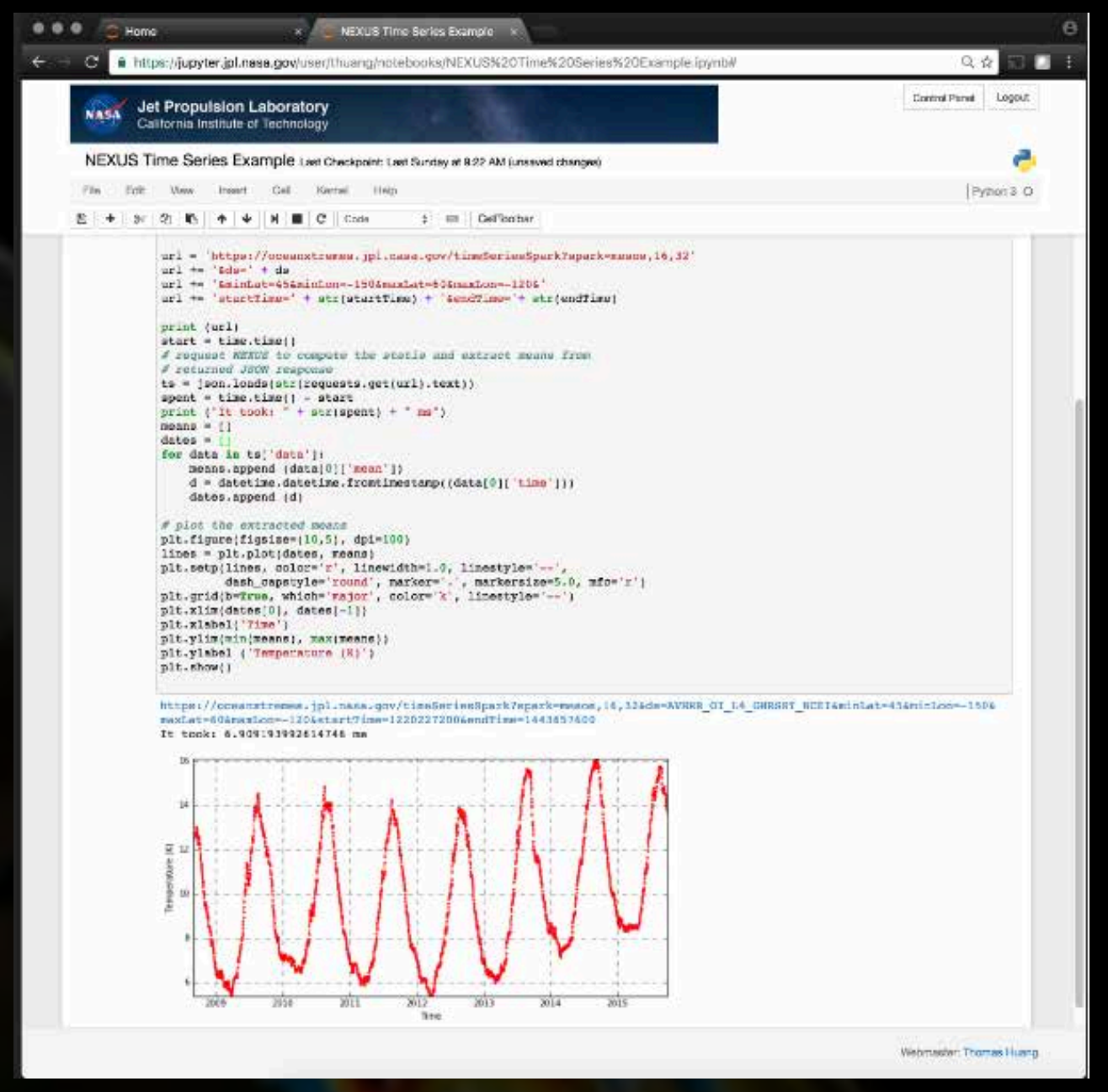
ABSTRACT

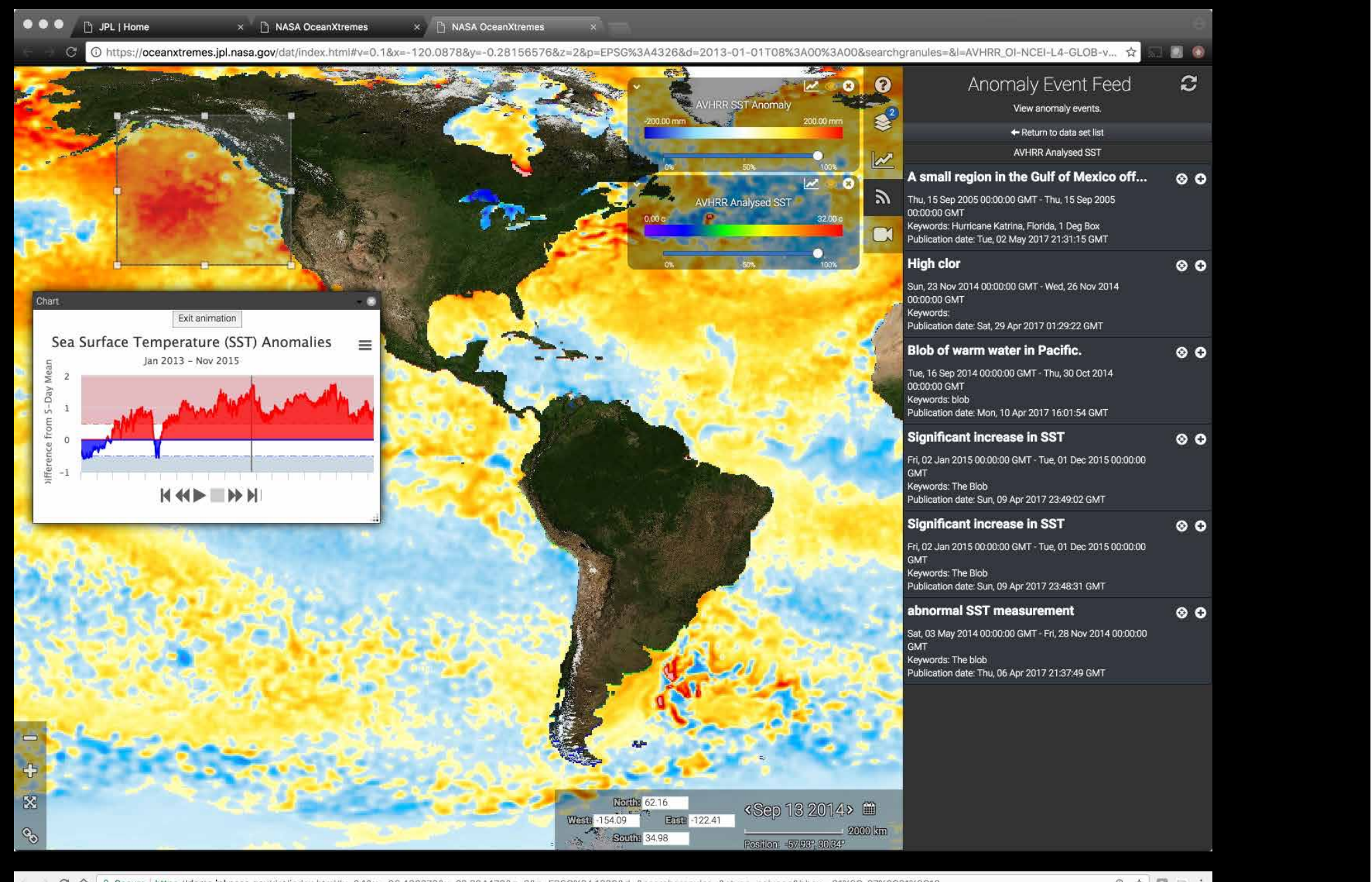
The Apache **Science Data Analytics Platform (SDAP) (http://sdap.incubator.apache.org)** is a suite of Big Data solutions created through the support of the NASA Advanced Information Technology (AIST) and the NASA Earth Science Data System (ESDS) programs. SDAP is led by the NASA Jet Propulsion Laboratory through collaboration with George Mason University, the Center for Ocean-Atmospheric Prediction Studies (COAPS) at the Florida State University, and the National Center for Atmospheric Research (NCAR). Its goal is to create a community-supported, integrated platform for big geospatial data analysis using Cloud computing technology. The SDAP currently includes NEXUS, OceanXtremes, DOMS, EDGE, and MUDROD.

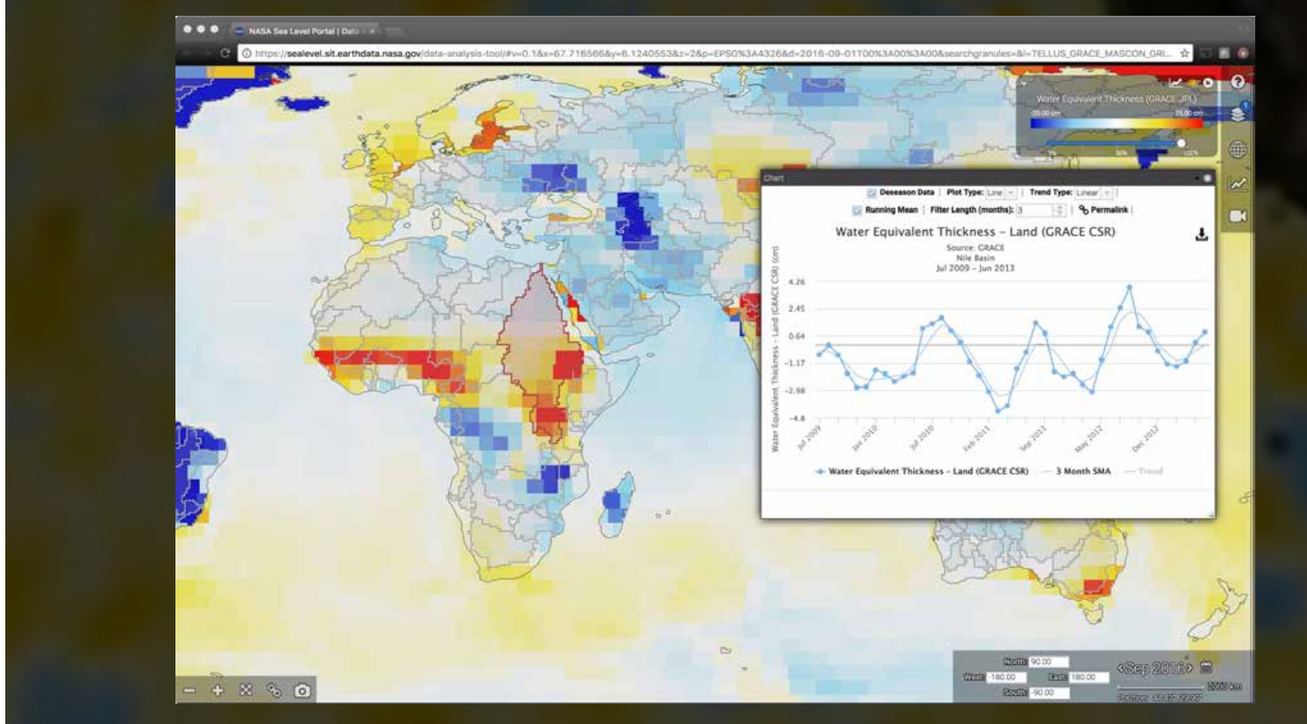
- **NEXUS** provides a suite of on-the-fly data analysis services including time series generation, area averaged map, climatological map, etc. that are essential to climate research. It can analyze data hundreds of times faster than traditional file-based analysis method. All the NEXUS' analytic capabilities are exposed as RESTful API, hiding the complexity of horizontal-scaling and map-reduced computing.
- **OceanXtremes** is a data-intensive anomaly detection solution that is built on the NEXUS solution. It provides cloud-based climatology generation and on-the-fly comparison of observational data against the climatology. OceanXtremes is equipped with the ability for researcher to document, share, and re-create identified ocean anomalies.
- **Distributed satellite and in situ matchup** – The Distributed Oceanographic Matchup Service (DOMS) delivers a cloud-based matchup solution by integrating distributed in situ data hosted at JPL, NCAR, and COAPS. The project has standardized access to point-based in situ data using open source implementation of OpenSearch called the Extensible Data Gateway Environment (EDGE). DOMS translates the temporal spatial query into in situ subset requires to the external data centers. Upon receiving the subsetting in situ data, DOMS executes its map-reduced, matchup algorithm on the cloud. The matchup result is packed in CSV/netCDF and visualized.
- **Extensible Data Gateway Environment** is an implementation of the standard OpenSearch specification (http://www.opensearch.org) using Apache Solr or ElasticSearch as the backend repository. Using this integration platform, data provider and expose their holdings using standard OpenSearch, RESTful API. The technology has already been infused in several production environments.
- **Mining and Utilizing Dataset Relevancy from Oceanographic Dataset (MUDROD)** is a search analytic technology by continuously mining search logs from data portals. Through machine learning technology, MUDROD exposes hidden relationships between ocean datasets and dynamically adjust data ranking to show the most relevant datasets first.

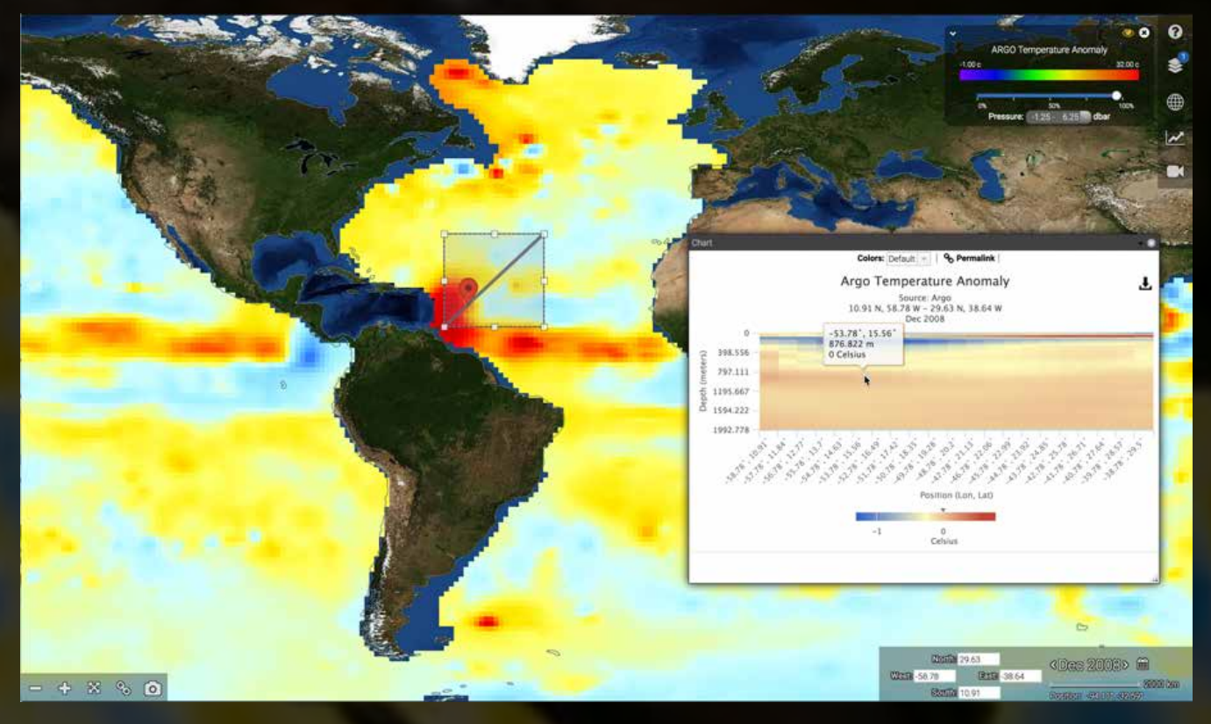
APPLICATIONS

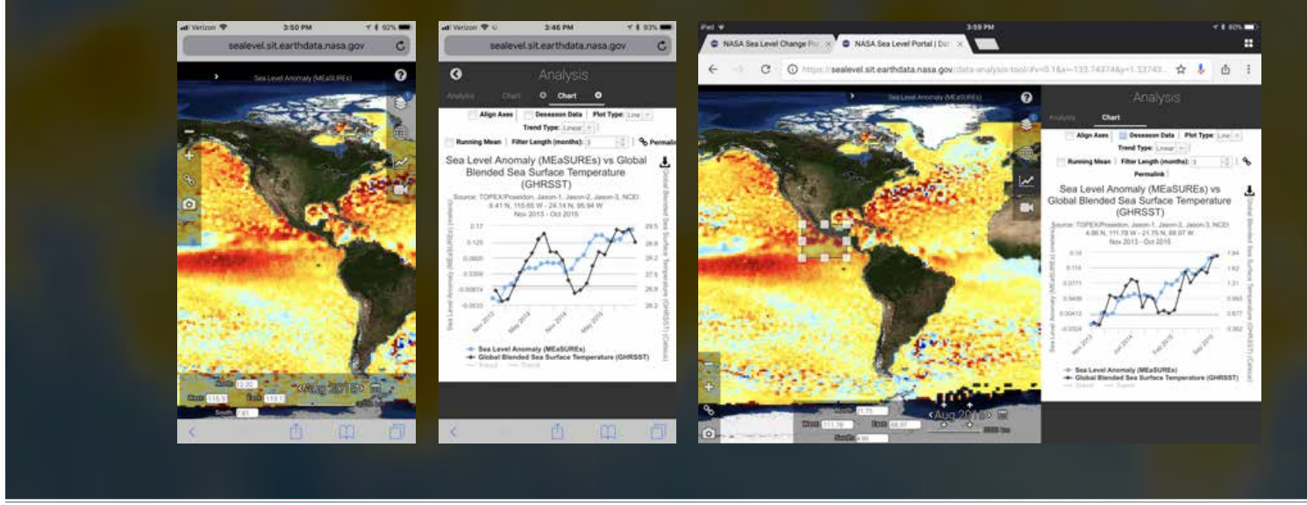


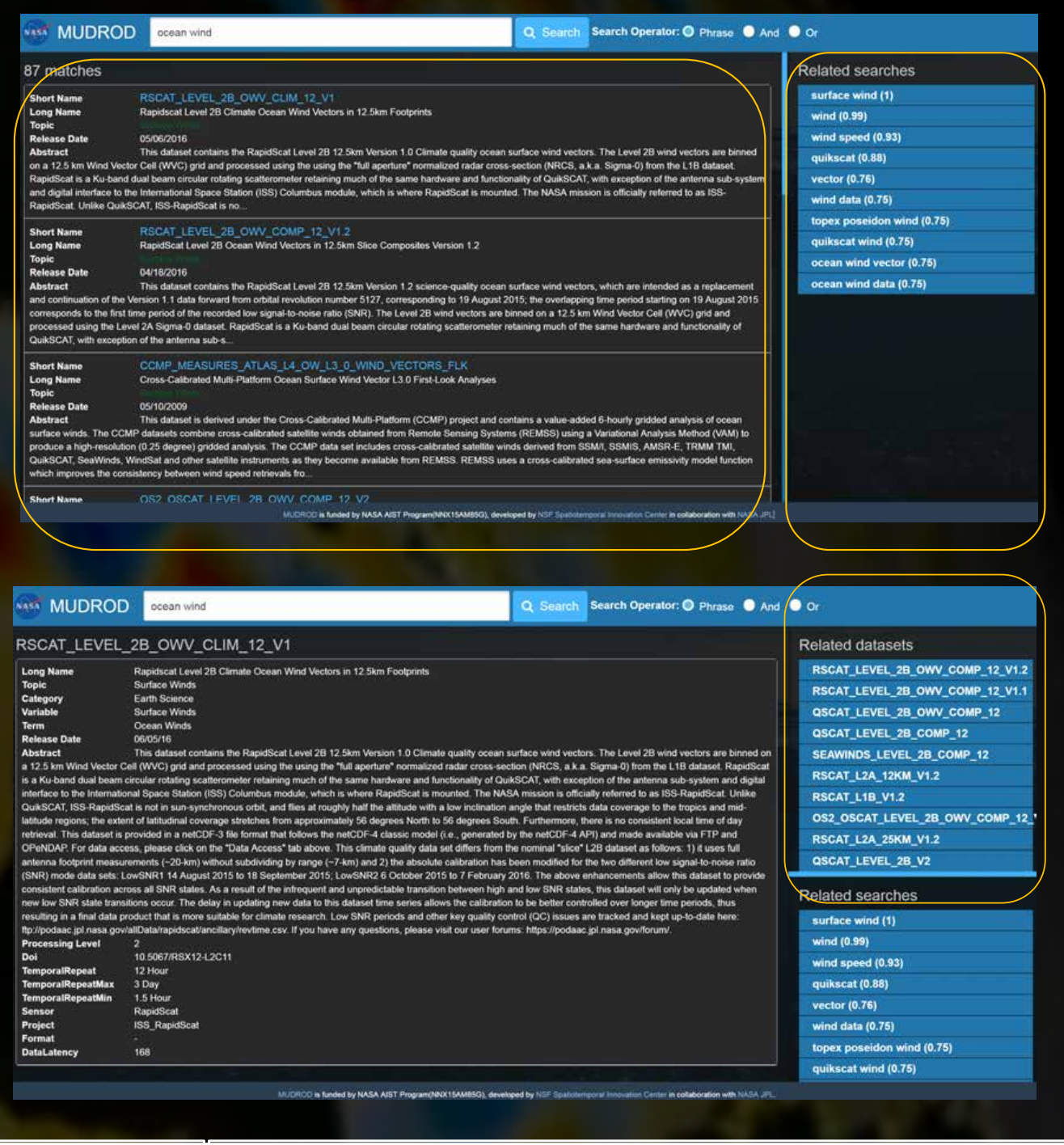


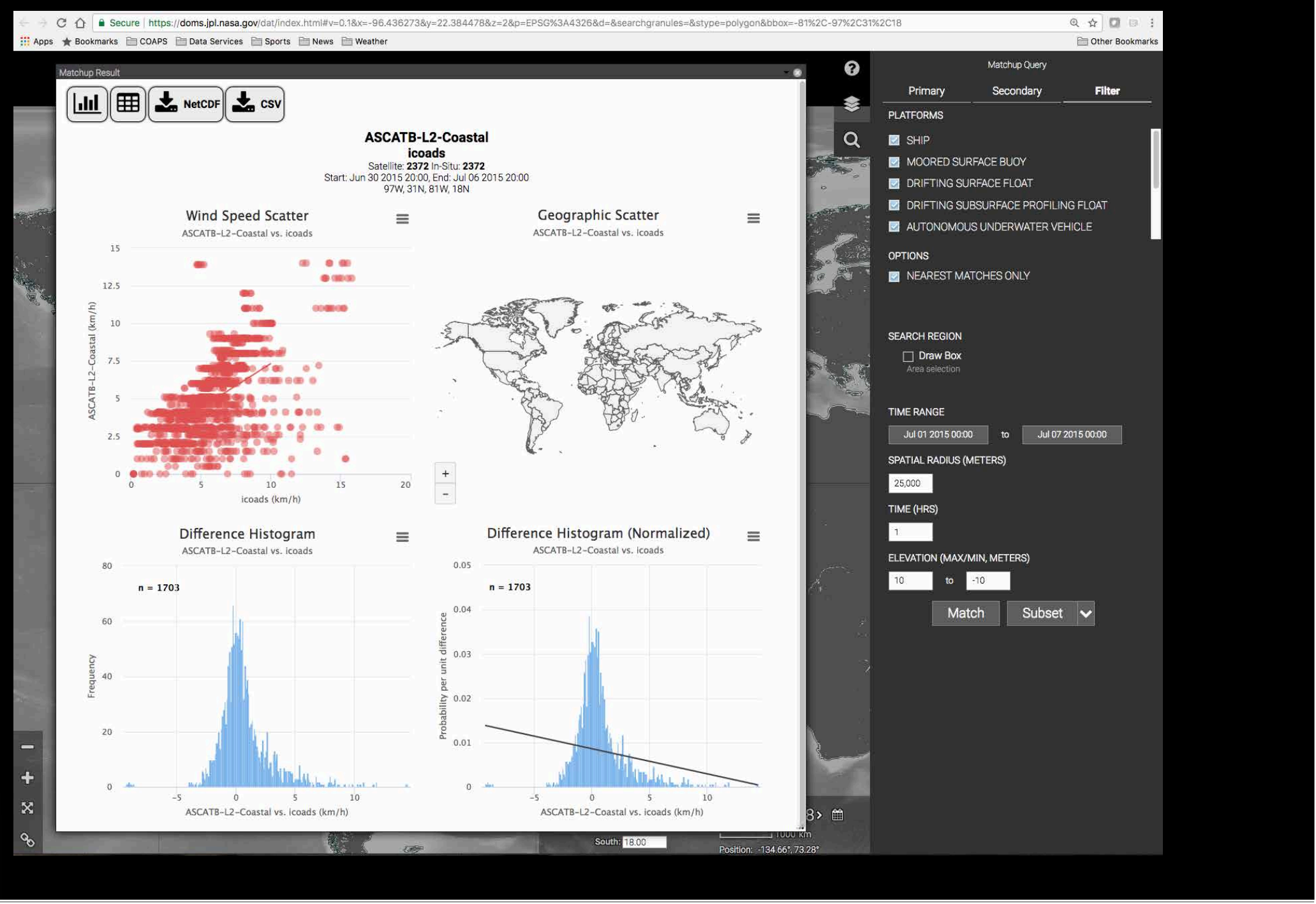






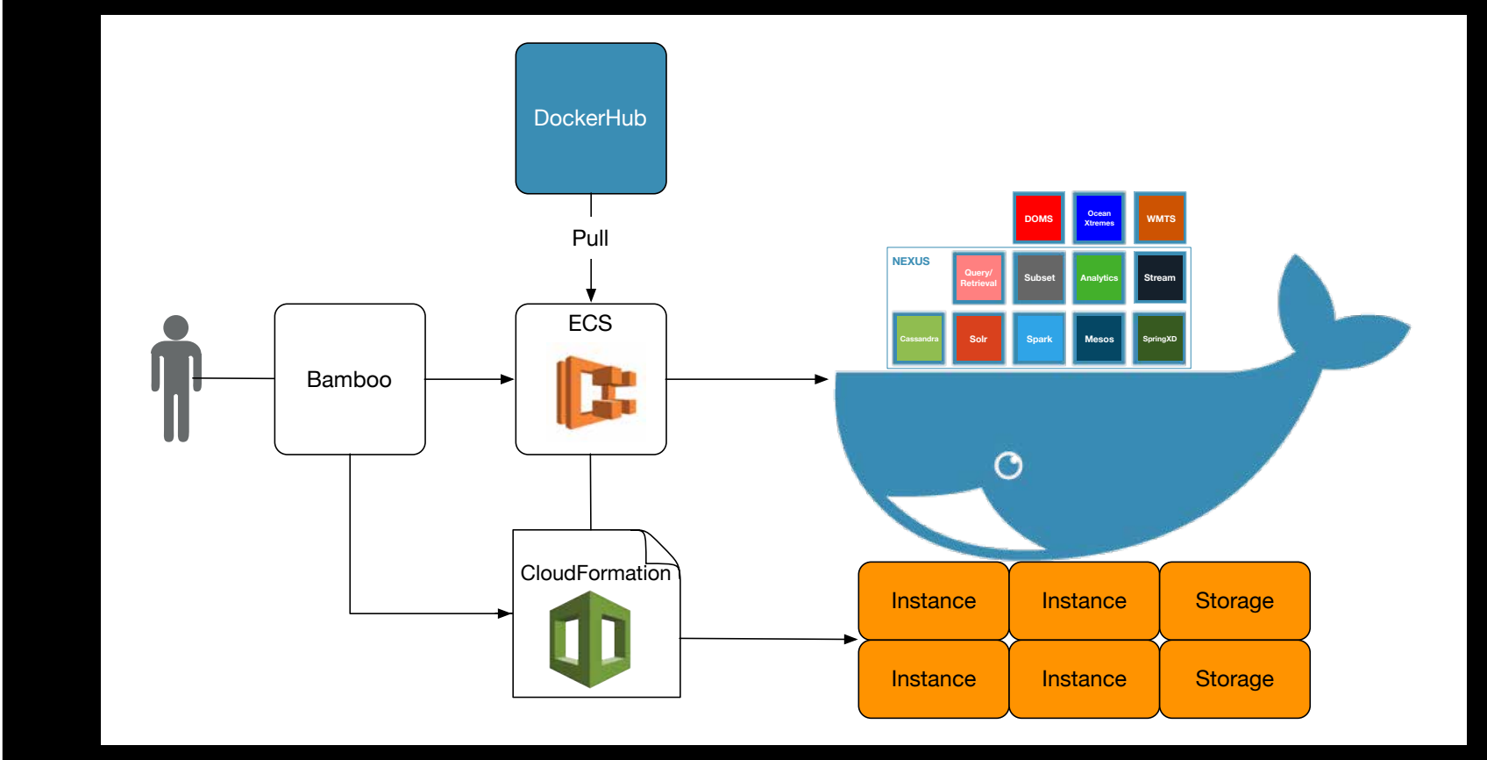


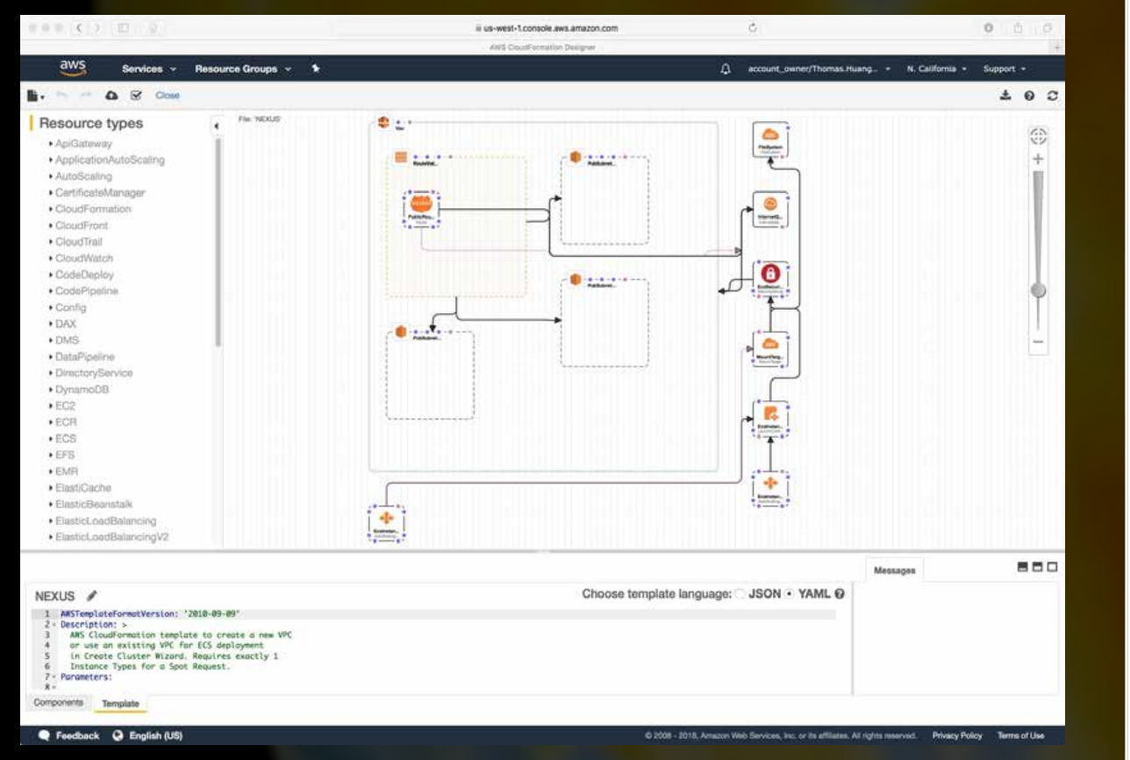




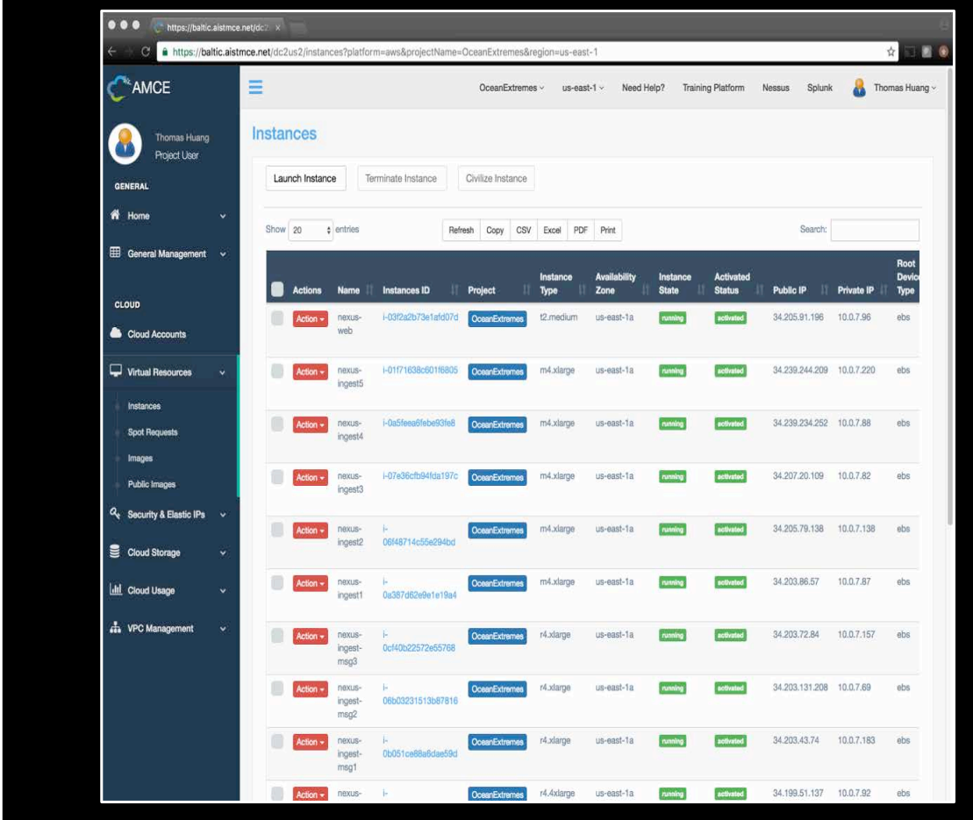
https://sealevel.nasa.gov
https://oceanxtremes.jpl.nasa.gov
https://doms.jpl.nasa.gov
https://mudrod.jpl.nasa.gov
https://jupyter.jpl.nasa.gov

CLOUD DEPLOYMENT AUTOMATION and CONTAINER-BASED

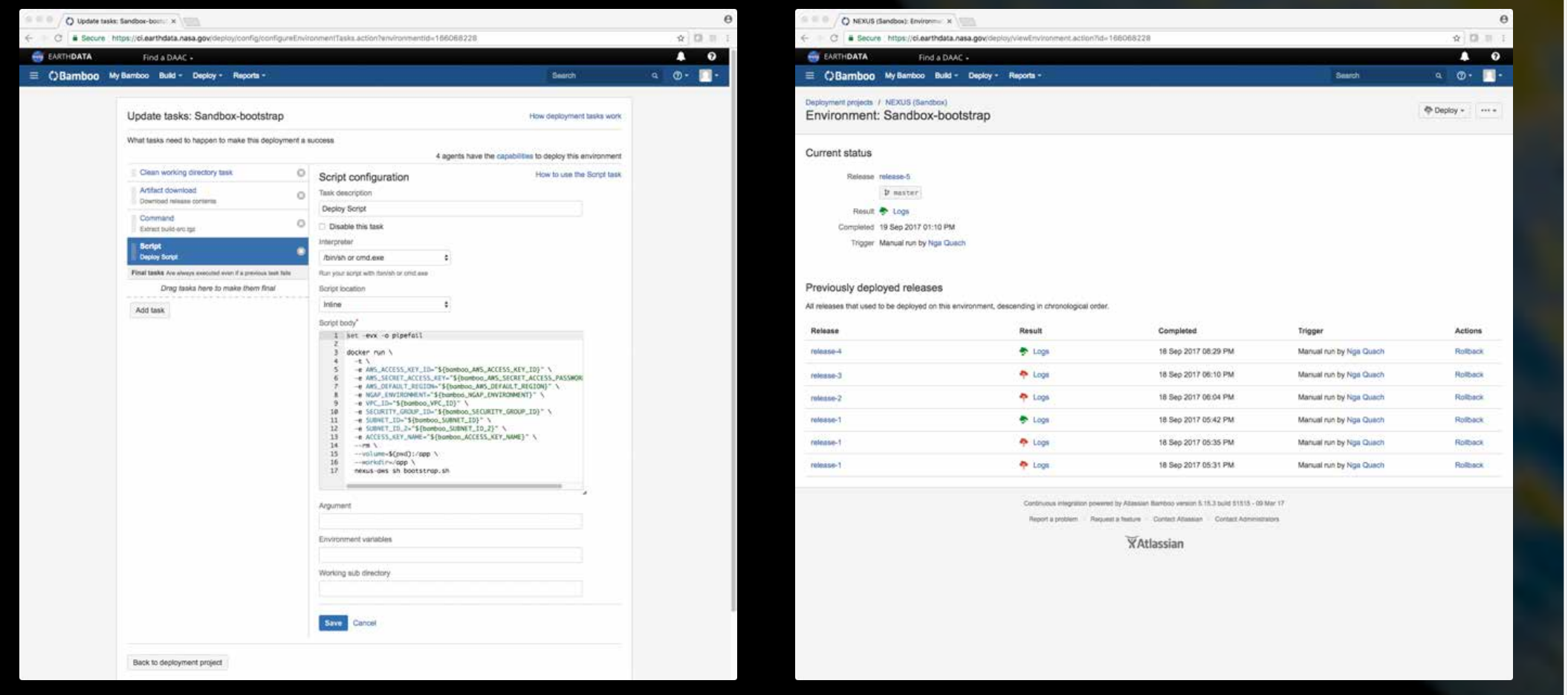




CloudFormation - INFRASTRUCTURE AUTOMATION



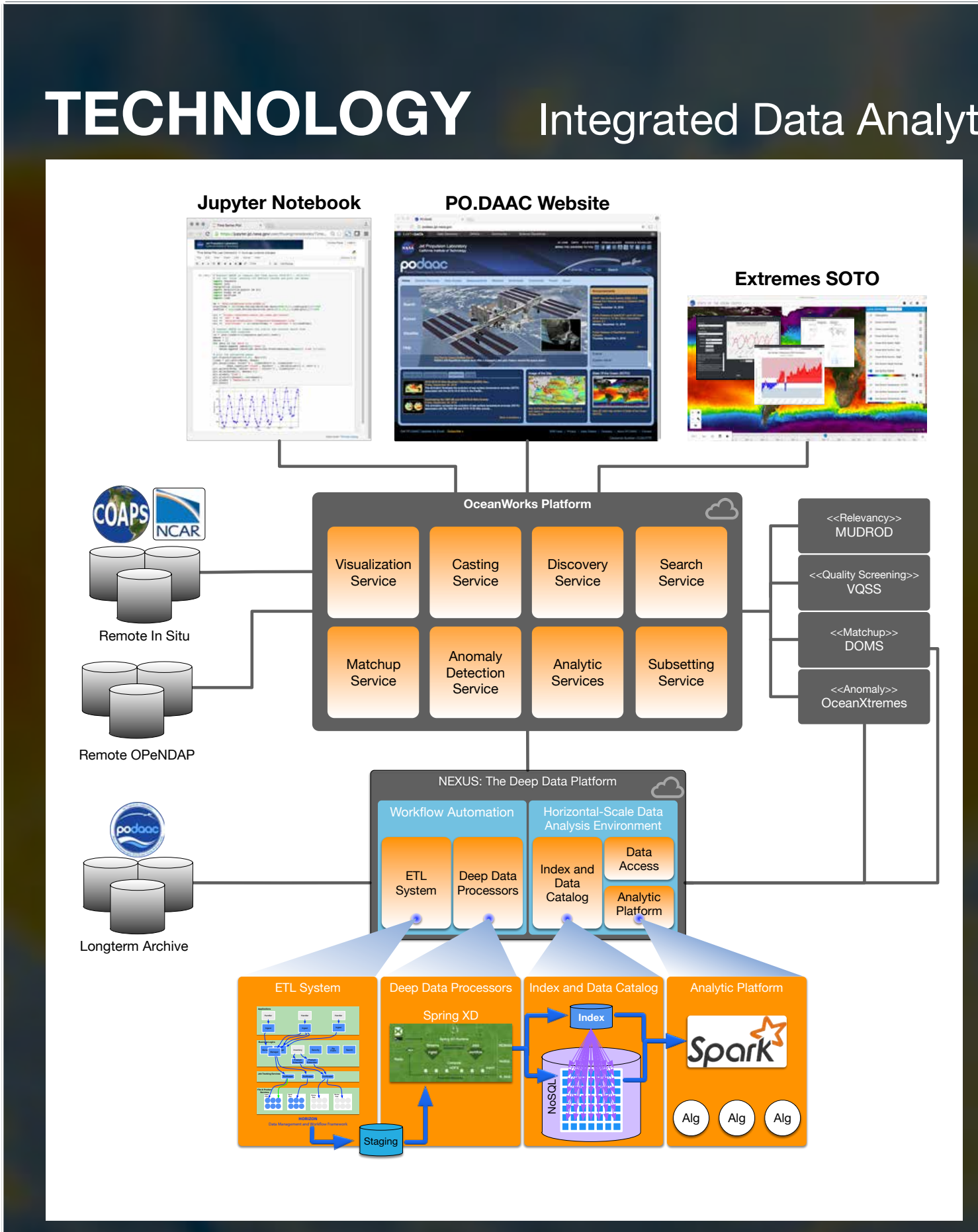
NASA AIST Managed Cloud Environment



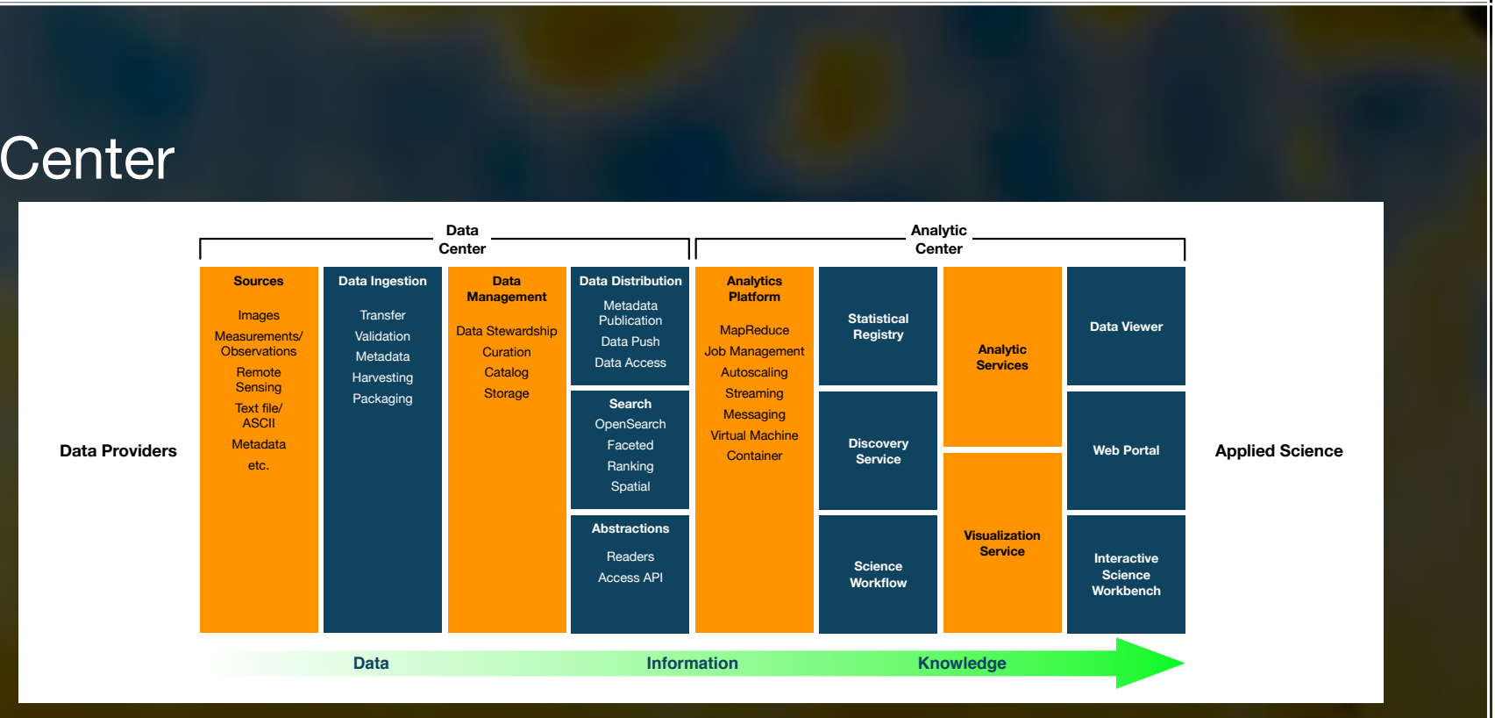
NASA Next Generation Application Platform (NGAP)

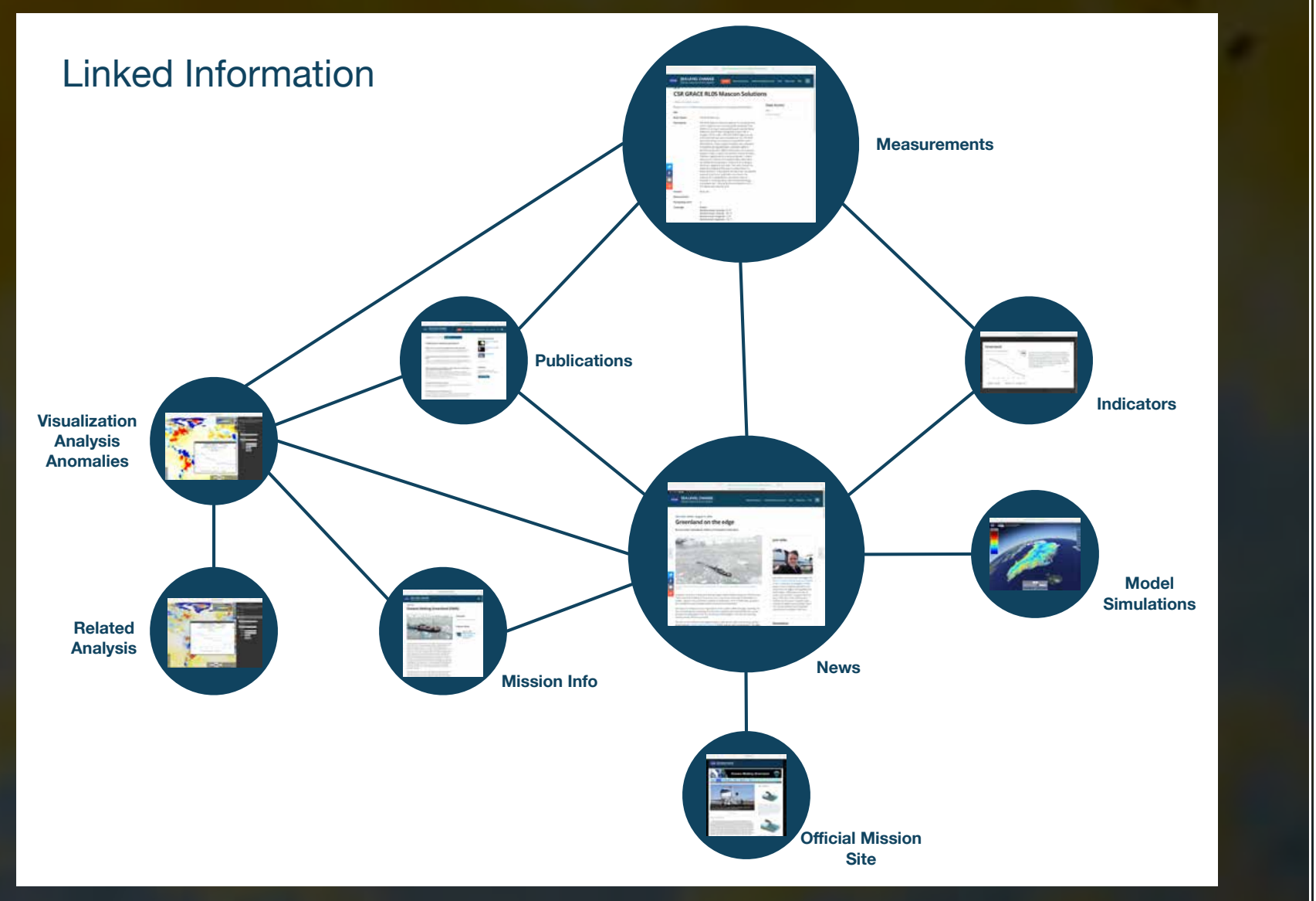
AWS AMCE MIRANTIS openstack

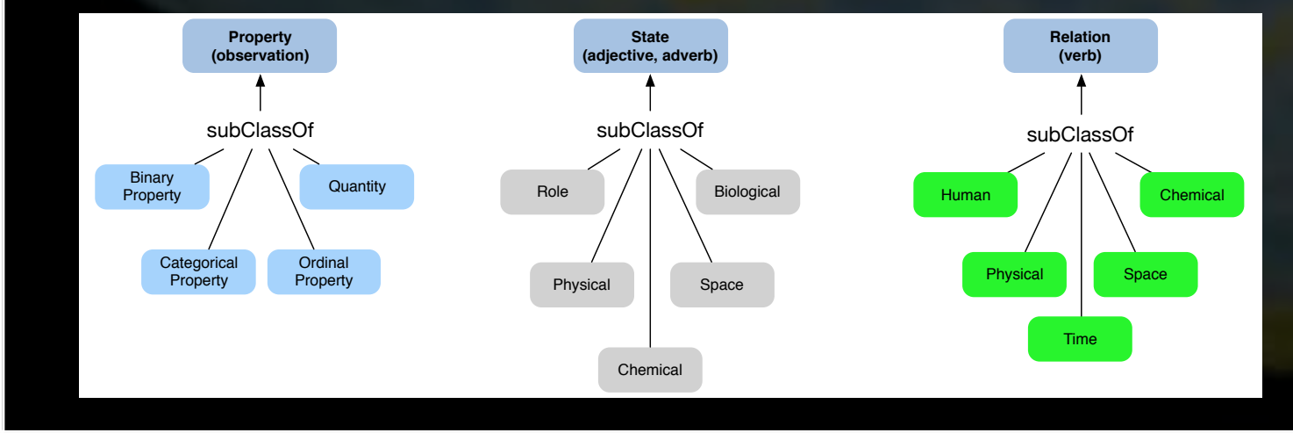
TECHNOLOGY Integrated Data Analytic Center

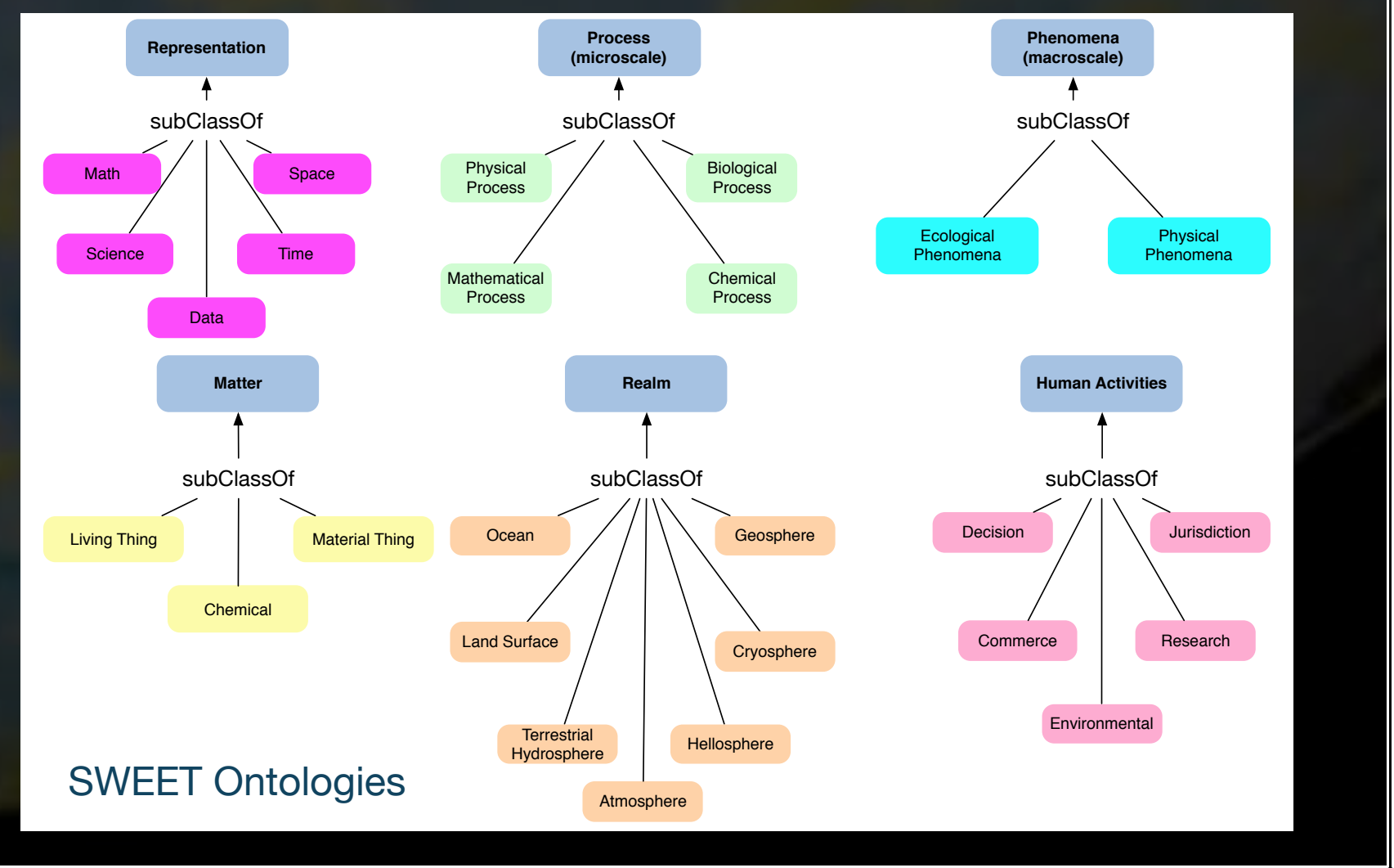


OceanWorks is an AIST Adjunct project to establish an Integrated Data Analytic Center at the NASA PO.DAAC for Big Ocean Science. It focuses on technology integration, advancement and maturity by bringing together several previous NASA-funded AIST and ACCESS projects as an effort to deliver a production-ready data science platform for the ocean science community. While its target is the ocean science community, the building blocks of OceanWorks are designed to support multidiscipline Earth Science. The emphasis is integration and platform building by hiding all the complexities of data management, domain-specific technology implementations, and cloud computing architecture. User applications and services will integrate with OceanWorks through RESTful APIs and well-defined information model. OceanWorks is a collaborative development effort between JPL, Center for Atmospheric Prediction Studies (COAPS) at Florida State University, National Center for Atmospheric Research (NCAR), and George Mason University (GMU).









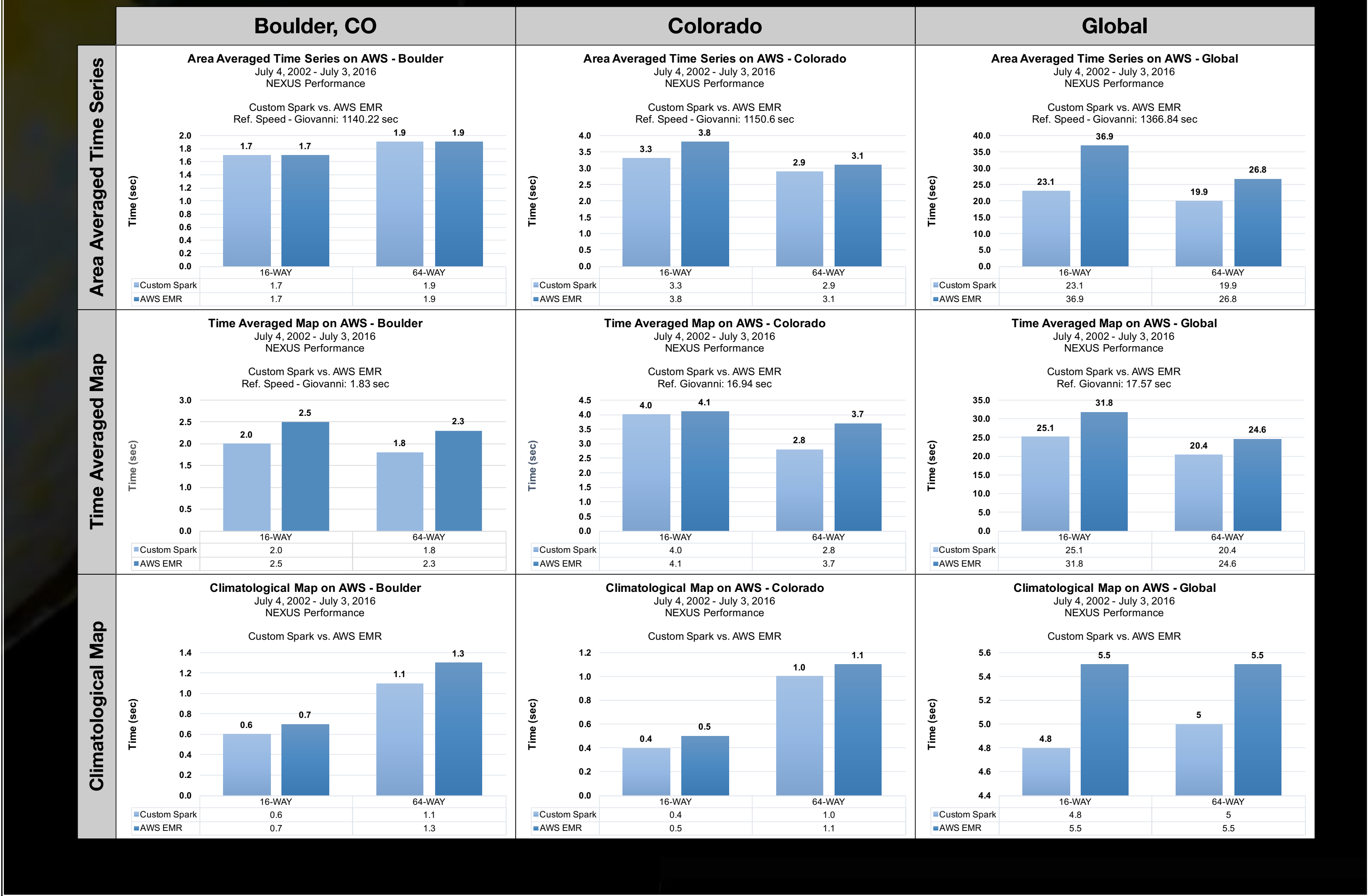
SWEET Ontologies

PERFORMANCE NEXUS (CUSTOM SPARK+MESOS) vs. EMR vs. GIOVANNI

Dataset Selection
Name: MODIS Aqua Daily L3 Atmospheres, Collection 6, Aerosol Optical Depth 550 nm (Dark Target) (MYD08_D3v6)
Duration: July 4, 2002 – July 3, 2016
File Count: 5106
Volume: 2.6GB

AWS Instances

	AWS Instance	Quantity	Cost
Custom Spark	r4.8xlarge	1	$2.128/hr
AWS EMR	r4.8xlarge	2 (1 EMR, 1 Data Mgmt.)	2 * $2.128/hr + $0.27/hr (EMR) = $4.526/hr



Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsements by the United States Government or the Jet Propulsion Laboratory, California Institute of Technology.

APACHE INCUBATOR